Model-Based Detection of Whole-Genome Duplications in a Phylogeny

Arthur Zwaenepoel*,[1,2,3] and Yves Van de Peer*,[1,2,3,4]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[2]Center for Plant Systems Biology, VIB, Ghent, Belgium

[3]Bioinformatics Institute Ghent, Ghent, Belgium

[4]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa

***Corresponding authors:** E-mails: arzwa@psb.vib-ugent.be; yvpee@psb.vib-ugent.be.

Associate editor: Jianzhi Zhang

Abstract

Ancient whole-genome duplications (WGDs) leave signatures in comparative genomic data sets that can be harnessed to detect these events of presumed evolutionary importance. Current statistical approaches for the detection of ancient WGDs in a phylogenetic context have two main drawbacks. The first is that unwarranted restrictive assumptions on the "background" gene duplication and loss rates make inferences unreliable in the face of model violations. The second is that most methods can only be used to examine a limited set of a priori selected WGD hypotheses and cannot be used to discover WGDs in a phylogeny. In this study, we develop an approach for WGD inference using gene count data that seeks to overcome both issues. We employ a phylogenetic birth–death model that includes WGD in a flexible hierarchical Bayesian approach and use reversible-jump Markov chain Monte Carlo to perform Bayesian inference of branch-specific duplication, loss, and WGD retention rates across the space of WGD configurations. We evaluate the proposed method using simulations, apply it to data sets from flowering plants, and discuss the statistical intricacies of model-based WGD inference.

***Key words:* genome duplication, phylogenetics, gene content evolution, Bayesian inference, reversible-jump MCMC.**

Introduction

"Every gene from a pre-existing gene," proclaimed Muller, echoing Virchow's famous third dictum (Muller 1936). This principle, though no longer thought of as a rule without exception (Long et al. 2013; Ruiz-Orera et al. 2018; Zhang et al. 2019), is still the paradigm for thinking about gene content evolution, with gene duplication regarded as the primary driver. A diverse array of molecular processes causes gene duplication (and loss) (Lynch 2007), and it is generally insightful to distinguish two main classes. Small-scale duplication and loss (SSDL) events, resulting in duplication or loss of a single or couple of genes, are thought to originate from processes such as nonhomologous recombination and transposition. These may be contrasted with large-scale duplication and loss events involving the whole genome or a large fraction of it, which generally result from chromosome-level processes such as aneuploidy and rediploidization after polyploidization. We refer to the latter process by the colloquial term *whole-genome duplication* (WGD). It is nowadays well appreciated that ancestral WGD has been of considerable importance in the evolution of eukaryotic genomes (Van de Peer et al. 2017), and especially so in plants, where it would be safe to state that the genome of every extant plant has been shaped to some degree by ancient WGD events (1KP initiative 2019). Identifying ancient WGDs through comparative genomic analyses remains however a nontrivial task, as exemplified by several recent controversies (Li et al. 2019; Nakatani and McLysaght 2019; Zwaenepoel et al. 2019), illustrating the need for reliable statistical methods to detect WGDs in a phylogenetic context.

One of the consequences of Muller's adage is that a particular class of stochastic models, namely birth–death processes (BDPs), has since long been a mainstay for modeling the evolution of gene family sizes (Novozhilov et al. 2006). In particular, the linear BDP and its variants have been widely used because of their tractable transition probabilities, facilitating statistical inference in a phylogenetic context (Hahn et al. 2005; Csűrös and Miklós 2009; Liu et al. 2011; Librado et al. 2012; Rabier et al. 2014; Tiley et al. 2016; Tasdighian et al. 2017). In these approaches, a series of linear BDPs are assumed to operate along the branches of a known species tree, providing an intuitive generative model for gene counts in a set of taxa for a single gene family. Such a set of gene counts is often referred to as a *phylogenetic profile* and can be obtained from genomic data sets using standard comparative genomics methods, thereby providing an attractive data set for likelihood-based inference of phylogenetic BDP models. As a model of gene family evolution, the birth and death rate parameters of the linear BDP are generally interpreted as the per-gene rate of duplication and loss per unit of time, thereby constituting a reasonable model of gene family evolution by SSDL (Novozhilov et al. 2006). Interestingly, large-scale duplication and loss, and in particular WGD, cause a

doi:10.1093/molbev/msaa111 Advance Access publication May 2, 2020

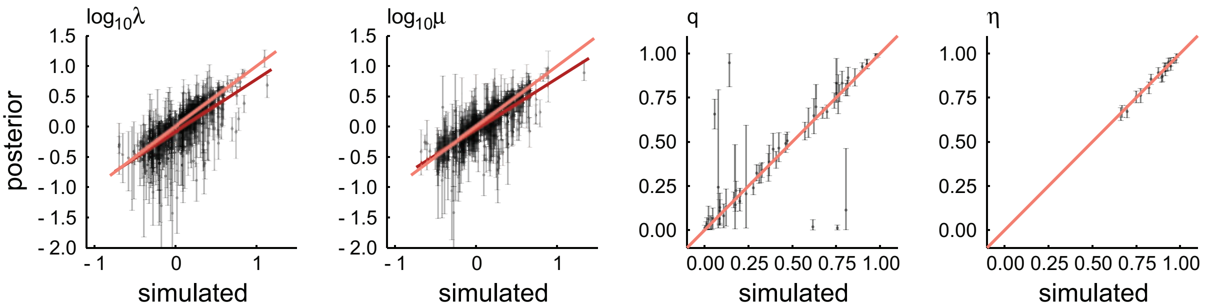

Fig. 1. Comparison of posterior means and 95% uncertainty intervals with simulated values. Results across different simulated data sets are pooled together in one plot. In orange, the ideal (expected) relation between posterior means and true (simulated) parameter values is shown, whereas in red, the least squares regression of the posterior means on the true values is shown. We refer the reader to supplementary figure S7, Supplementary Material online, for plots of the retention rate estimates by replicate.

characteristic deviation from the SSDL-driven evolutionary process. Indeed, posterior predictive simulations (performed using methods and data discussed below) clearly show that WGDs that are not accounted for are the main source of lack of fit of the phylogenetic linear BDP in a plant data set with several well-known WGD events (supplementary figs. S1–S3, Supplementary Material online). Furthermore, these unmodeled WGDs cause biases in the estimated duplication and loss rates, compromising their interpretation as rates of the SSDL process.

Rabier et al. (2014) were the first to propose a model of gene family evolution including ancestral WGDs, exploiting the deviation from a linear BDP to infer WGDs in a statistically founded approach. The same model was adopted in Zwaenepoel and Van de Peer (2019) in a gene tree reconciliation based inference context. In that study, we showed that across-lineage variation in gene duplication and loss rates is a crucial factor that should be taken into account when performing model-based WGD inference, and we developed a Bayesian approach modeling the variation in duplication and loss rates using relaxed molecular clock priors. In both Rabier et al. (2014) and Zwaenepoel and Van de Peer (2019), standard model selection techniques are used to determine whether some WGD hypothesis provides a significant better fit to the observed gene trees or phylogenetic profiles. These methods are therefore restricted to testing a limited set of a priori determined WGD hypotheses and are unable to "automatically" detect WGDs from the data. In this study, we develop an approach based on reversible-jump Markov chain Monte Carlo (rjMCMC) to infer WGDs under a probabilistic model of gene family evolution in a phylogenetic context based on a set of phylogenetic profiles without the need of specifying a restricted set of a priori WGD hypotheses. We study the performance of the new approach using simulations and explore its practical utility by applying the method to several comparative genomic data sets from plants.

New Approaches

We implement a Bayesian inference approach based on reversible-jump MCMC to perform model-based detection of ancient WGDs and branch-specific duplication and loss rates for a given species phylogeny based on a set of phylogenetic profiles. Our approach does not assume an a priori determined set of putative WGDs as in Rabier et al. (2014) or Zwaenepoel and Van de Peer (2019). Our method is based on the DLWGD model of gene family evolution, combining a phylogenetic BDP—formulated as in Csűrös and Miklós (2009)—as a model for the SSDL process, with a model for WGD as in Rabier et al. (2014). We employ a flexible hierarchical model for the gene family evolutionary process, modeling the evolution of gene duplication and loss rates using a bivariate stochastic process operating along the species tree, in a vein similar to Lartillot and Poujol (2011).

Materials and Methods

Probabilistic Model of Gene Content Evolution

We employ a probabilistic model of gene family evolution based on the linear BDP to model the evolution of the number of genes in a gene family along a time-calibrated species tree $\mathcal{S}$ with $V(\mathcal{S})$ nodes (Hahn et al. 2005; Csűrös and Miklós 2009; Liu et al. 2011; Librado et al. 2012; Rabier et al. 2014; Tasdighian et al. 2017). Specifically, we assume an independent linear BDP for each branch i of $\mathcal{S}$ with duplication and loss rates λ_i and μ_i. We adopt the WGD model of Rabier et al. (2014) to introduce WGDs in the phylogenetic BDP, and call the resulting model the DLWGD model (for duplication, loss, and WGD). Under the DLWGD model, we assume a set of k WGDs are indicated along $\mathcal{S}$, each characterized by a retention rate q_j and age t_j, $j = 1, \dots, k$. We denote by $b(j)$ the branch of $\mathcal{S}$ along which the WGD with index j is located. A set of k WGDs with the respective branches of $\mathcal{S}$ on which they are located will be called a "WGD configuration." At the time of a WGD, all extant gene lineages are duplicated, and a duplicated gene is either retained with probability q_j or not retained with probability $1 - q_j$. Under this model, the complex polyploidization–rediploidization process associated with a WGD is assumed to happen in a time interval that is short relative to the branch length on which the WGD occurs, such that rediploidization is effectively instantaneous compared with the time scale of the phylogeny and, crucially, has completed before the next speciation event. In other words, the number of genes retained in duplicate after rediploidization is modeled as a binomial random variable with "success" probability q_j. The resulting probabilistic graphical model is a straightforward generative model for the evolution of gene trees (and as a consequence, also gene counts) evolving by means of duplication, loss, and WGD with respect to an assumed species tree.

Parameter inference is based on a data set X consisting of N phylogenetic profiles, that is, observed gene counts at the leaves of $\mathcal{S}$. A single phylogenetic profile for family i is denoted as $X^{(i)} = \{X_j^{(i)} : j \in \text{leaves}(\mathcal{S})\}$, where we reserve subscripts for node indices and superscripts for family indices. We assume all $X^{(i)}$, $i = \{1, \ldots, N\}$ are independent and identically distributed (iid) under the same DLWGD model. The linear BDP has a countably infinite state space, making direct computation of the likelihood under the DLWGD model using the pruning algorithm not possible. In most previous studies (Hahn et al. 2005; Librado et al. 2012; Tasdighian et al. 2017), this issue was circumvented by truncating the transition probability matrix to some carefully but arbitrarily chosen bound, where a tradeoff exists between computational cost and accuracy. Alternatively, in a Bayesian approach one could augment the parameter space by the states at the internal nodes of $\mathcal{S}$ for every gene family, and sample from the augmented probability distribution using an MCMC algorithm, as was done by Liu et al. (2011). Following Rabier et al. (2014), we compute the likelihood for the resulting probabilistic graphical model using the algorithm of Csűrös and Miklós (2009), which uses *conditional survival likelihoods* in combination with a discrete prior distribution on the number of gene lineages at the root of a gene family to compute the marginal likelihood of a phylogenetic profile conditional on the assumed prior at the root. We employ a geometric prior on the number of lineages at the root with expected value $1/\eta$ following Rabier et al. (2014). For more details on the algorithm to compute the conditional survival likelihood under the DL and DLWGD model, we refer the reader to Csűrös and Miklós (2009) and Rabier et al. (2014).

Prior Distributions and Posterior Inference

In previous work, we showed that modeling across-lineage variation in duplication and loss rates is crucial for the assessment of WGD hypotheses (Zwaenepoel and Van de Peer 2019). In that study, we modeled duplication and loss rates using a stochastic molecular clock by adopting priors similar to those used in Bayesian divergence time estimation. However, there we assumed the evolution of duplication and loss rates to be two independent processes. This assumption may be innocuous but is not biologically plausible. Given that the expected value $\mathbb{E}[Y]$ of the linear BDP over a time t for a given initial state Y_0 is $Y_0 e^{(\lambda-\mu)t}$; for a gene family not to systematically expand or contract, λ should be approximately equal to μ for each branch in $\mathcal{S}$ under the assumed model. A strong correlation between λ and μ is therefore expected a priori under what can be called a scenario of "neutral" gene family evolution.

In the present work, we draw inspiration from Lartillot and Poujol (2011) and propose two bivariate models for across-lineage rate variation. The first model, faithful to Lartillot and Poujol (2011), models the evolution of λ and μ by a bivariate Brownian motion (BM) with some unknown covariance matrix Σ. The multivariate BM specifies a conditional probability density on the state of a random vector $\theta_i = (\log \lambda_i, \log \mu_i)$ for each *node* i of $\mathcal{S}$, where we condition on the state of the parent node j at distance t_i. Specifically, we have that

$$\theta_i | \theta_j, t_i \sim \text{Normal}(\theta_j, \Sigma t_i).$$

To obtain branch-specific duplication and loss rates, we approximate the sample path of the BM along a branch of the tree by taking the arithmetic average of the rates at the two flanking nodes. Alternatively, we consider a second model where rates across branches of $\mathcal{S}$ are independent and identically distributed, that is

$$\theta_i | \theta_1 \sim \text{Normal}(\theta_1, \Sigma),$$

where we denote the branch leading to node i with the index i and denote for a branch with index i by θ_i the vector of the logarithm of the branch-specific duplication and loss rates. In contrast to the bivariate BM model, this bivariate IR model can be specified directly in terms of branch-specific rates, resulting in considerable improvements in the efficiency of the MCMC algorithms we develop (see further). For both the BM and IR model, we assume an inverse Wishart prior on the covariance matrix Σ with prior covariance matrix Ψ and degrees of freedom $q = 3$ (Lartillot and Poujol 2011). For the state at the root in the BM model, or the mean rates in the IR model (θ_1), we assume a multivariate Normal prior with covariance matrix Σ_0 (which may or may not be chosen identical to Ψ) and prior mean vector θ_0.

We denote the total tree length by T and assume a mapping from points in the interval $[0, T]$ to points along $\mathcal{S}$, similar to, for instance, Huelsenbeck et al. (2000). We assume a uniform prior on the interval $[0, T]$ for the WGD times and a Beta prior for the associated retention rates q. Note that we treat the WGD times as random variables, whereas in previous studies, these were usually assumed fixed, a potentially problematic assumption given the difficulties associated with dating uncertain WGD events (Rabier et al. 2014; Tiley et al. 2016; Zwaenepoel and Van de Peer 2019). Lastly, we consider a Beta prior on the η parameter that specifies the geometric prior on the number of lineages at the root. The full hierarchical model can be summarized as (taking the IR model as an example, see also supplementary fig. S4, Supplementary Material online)

$$\begin{aligned}
\eta &\sim \text{Beta}(a_\eta, b_\eta) \\
q_j &\sim \text{Beta}(a_q, b_q) \qquad j = 1, \ldots, k \\
t_j &\sim \text{Uniform}(0, T) \qquad j = 1, \ldots, k \\
\Sigma &\sim \text{InverseWishart}(\Psi, 3) \\
\theta_1 &\sim \text{Normal}(\theta_0, \Sigma_0) \\
\theta_i | \theta_1, \Sigma &\sim \text{Normal}(\theta_1, \Sigma) \qquad i = 2, \ldots, V(\mathcal{S}) \\
X | \theta, q, t, \eta &\sim \text{DLWGD}(\mathcal{S}, \theta, q, t, \eta)
\end{aligned} \quad .$$

Unless stated otherwise, we used the following "vague" prior settings: a Beta$(3, 1)$ prior for η, a Beta$(1, 3)$ prior for

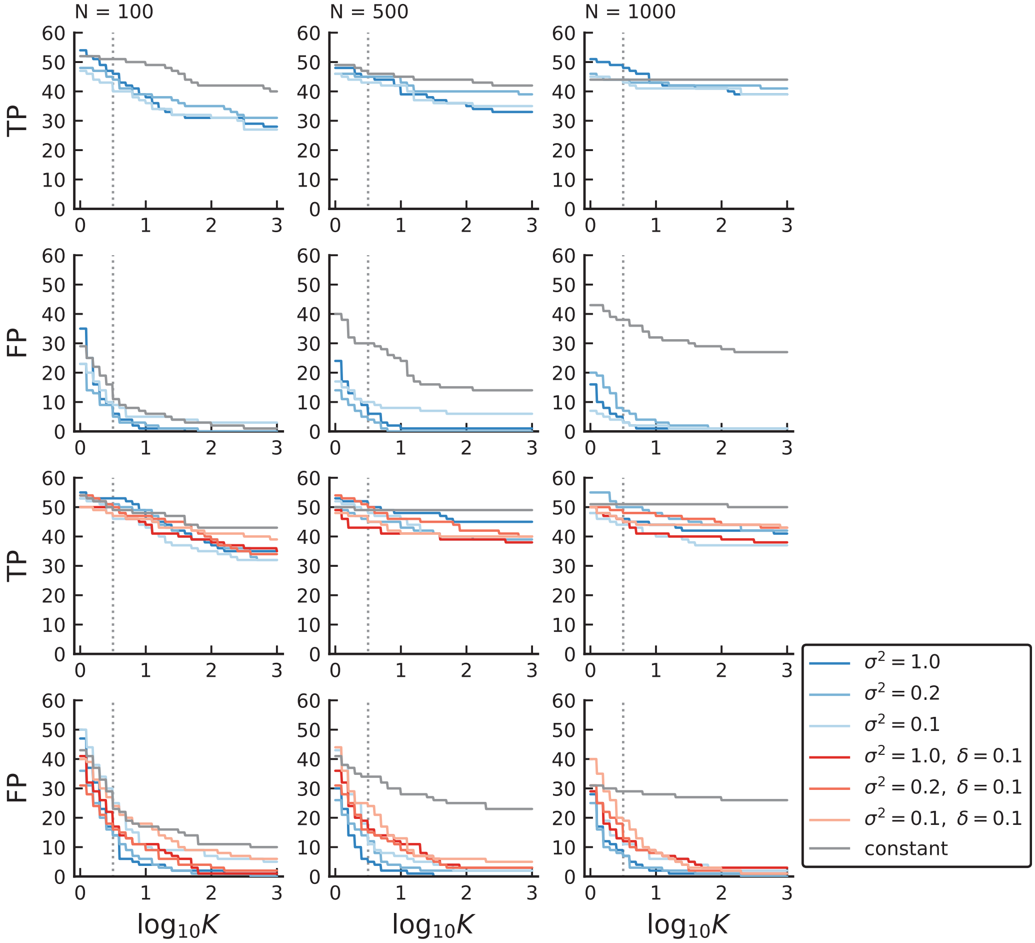

FIG. 2. Performance of the rjMCMC sampler to detect simulated WGDs in different simulated data sets. The number of true positive (TP) and false positive (FP) WGDs detected for increasing cutoffs of the Bayes factor ($\log_{10}K$) is shown. The top two rows show the simulations from the IR prior, whereas the bottom two rows show the simulations based on the joint posterior distribution for the dicots data set (see main text). In both series, we show results for data sets of size $N = 100$, $N = 500$, and $N = 1000$ and a prior covariance matrix of $\sigma^2 I_2$, with $\sigma^2 \in \{1, 0.2, 0.1\}$. We include results for the constant rates prior as well. Results are pooled across ten replicate simulations. The dotted line marks a rule-of-thumb threshold of 0.5 for the base 10 logarithm of the Bayes factor ($\log_{10}K$).

q, $\theta_0 = [\log(1.5),\ \log(1.5)]$, $\Sigma_0 = [1\,0.9\,;\,0.9\,1]$, and $\Psi = I_2$, that is, the 2×2 identity matrix. We make the important remark that any particular choice of prior parameterization for θ_0, Σ_0 and Ψ should depend on the time scale of the phylogeny. We sometimes consider an additional prior on the expected number of lineages at node i given one lineage at the parent node j, $\mathbb{E}[X_i|X_j = 1] = e^{(\lambda_i - \mu_i)t_i}$, which serves to constrain the duplication and loss rate to a regime where they are of similar magnitude. For this "constraining" prior we consider a Normal distribution with mean one and standard deviation δ. The advantage of this additional prior, which is not in a direct relation to the generative model, is that it allows to constrain duplication and loss rates in a biologically meaningful and insightful way due to its interpretation in terms of gene family expansion and contraction. Encoding such prior intuitions on gene family evolutionary dynamics in terms of the prior covariance matrix Ψ is generally less obvious.

We devised a Metropolis-within-Gibbs MCMC algorithm to obtain approximate samples from the posterior distribution for a model with a fixed WGD configuration. A single iteration of the MCMC algorithm involves a postorder traversal along the tree updating parameters using the following conditional updates:

$$\eta | \theta, q, t, \tag{1}$$

$$\theta_i | \theta_{-i}, \eta, q, t \qquad i = 1, \ldots, V(\mathcal{S}), \tag{2}$$

$$q_j, t_j | \theta, \eta, q_{-j}, t_{-j} \qquad j = 1, \ldots, k, \tag{3}$$

and

$$q_j, \theta_{b(j)} | \theta_{-b(j)}, \eta, q_{-j}, t \qquad j = 1, \ldots, k, \tag{4}$$

where we use a common notational convenience by denoting by θ_{-i} the vector θ with θ_i excluded. Note that we do not explicitly update the covariance matrix Σ but use the fact that the inverse Wishart distribution is a conjugate prior for the multivariate Normal to integrate out the covariance matrix Σ and compute $p(\theta|\Psi) = \int_\Sigma p(\theta|\Sigma)p(\Sigma|\Psi)\mathrm{d}\Sigma$ directly,

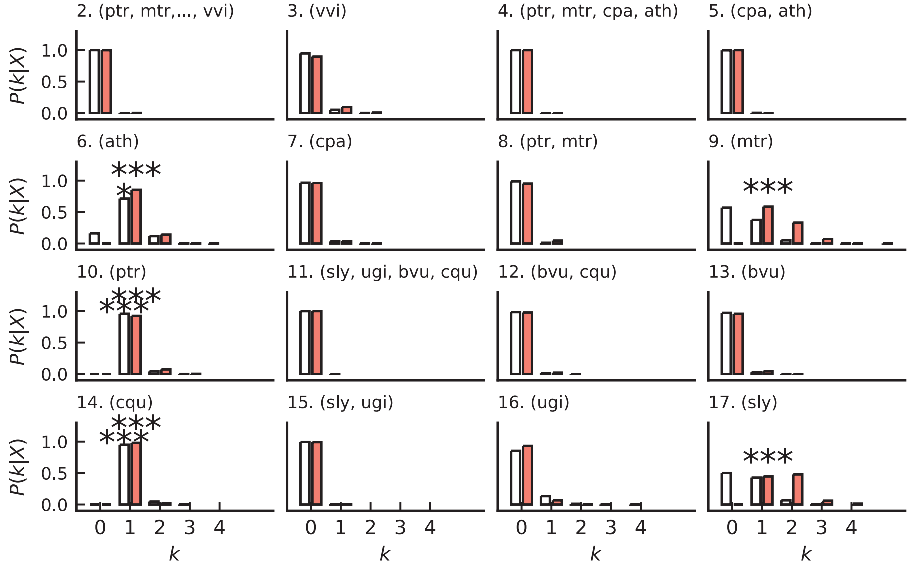

Fig. 3. Posterior probability of the number of WGDs on each branch in the dicot data set for an analysis with $\delta = 0.1$ (white) and an analysis with $\delta = 0.05$ (red). Asterisks indicate the magnitude of the associated Bayes factor (*) $0.5 < \log_{10} K < 1$, (**) $1 < \log_{10} K < 2$, and (***) $\log_{10} K > 2$. Three letter codes (see Materials and Methods) in between brackets denote the relevant clade below the particular branch.

as in Lartillot and Poujol (2011). For multivariate updates (2) and (4), we use a proposal strategy akin to Lartillot and Poujol (2011), where we propose a new state by executing one of the following moves uniformly at random: 1) additively update a uniformly chosen parameter using a uniform random walk proposal, 2) update all parameters by the same additive uniform random variable, and 3) update all parameters additively with iid uniform random variables. For update (3), we update either q_j, t_j or both each with probability one-third using a uniform random walk proposal for q_j and an independent proposal for t_j, sampled from the time interval between the parent and child node of WGD j. Lastly, we update η using a uniform random walk proposal. For both q and η updates, we reflect proposals that are out of bounds back into the (0, 1) interval. All proposal parameters were automatically tuned using diminishing adaptation during burn-in (Roberts and Rosenthal 2009).

Reversible-Jump MCMC

Previous work on the DLWGD model has considered the WGD hypotheses fixed and prespecified and resorted to standard model selection approaches using likelihood ratio tests or Bayes factors to select among a small set of nested candidate models or "WGD configurations" (Rabier et al. 2014; Zwaenepoel and Van de Peer 2019). Here, we treat the WGD configuration as unknown and seek to jointly perform inference of branch-wise duplication and loss rates, the number of WGDs k, and their locations and retention rates $\{t_j, q_j; j \in 1, \ldots, k\}$ from a set of exchangeable phylogenetic profiles $X = \{X^{(i)}\}_{i=1}^{N}$ from N gene families.

Different configurations of WGD hypotheses along a phylogeny constitute different instances of the DLWGD model with the dimensionality of the parameter space depending on the number of WGDs on the branches of $\mathcal{S}$. Denote by ϕ_k the parameters associated with the DLWGD model with branch-wise duplication and loss rates θ and k WGDs. By Bayes' theorem we have

$$p(\phi_k, k|X) = \frac{p(X|\phi_k, k)p(\phi_k|k)p(k)}{p(X)},$$

where $p(k)$ is a prior distribution on the number of WGDs. This prior can be any discrete univariate distribution (e.g., below we consider a discrete uniform distribution with domain from 0 to 20). The reversible-jump MCMC algorithm (Green 1995) allows to simulate a Markov chain with $p(\phi_k, k|X)$ as stationary distribution without requiring evaluation of the marginal likelihood $p(X)$. In the rjMCMC algorithm, we construct trans-dimensional moves that add and remove WGDs to the current state. In the formalism of Green (1995), we denote the current state of dimensionality n as ϕ, and propose a new state ϕ' of dimensionality $n' = n + 2$ (i.e., a forward move, adding a WGD) by drawing a vector of random numbers $u = (u_1, u_2, u_3)$ from a joint density g such that $(\phi', u') = h(\phi, u)$, where h is some deterministic function and u' are the random numbers from a joint density g' required for the reverse move from ϕ' to ϕ with the inverse function h' of h. Assume we introduce a WGD j on branch $b(j) = i$, we write $\phi = (\phi^{*}, \log \lambda_i)$ and construct our forward move such that $(\phi', u') = h(\phi, u) = h[(\phi^{*}, \log \lambda_i), (u_1, u_2, u_3)] = [(\phi^{*}, \log \lambda_i - u_1, u_2, u_3), u'_1] = [(\phi^{*}, \log \lambda'_i, q_j, t_j), u'_1]$. The acceptance probability for the forward move required for obtaining detailed balance is (Green 1995)

$$\alpha[(\phi, k), (\phi', k')] = \min\left\{1, \frac{p(X|\phi', k')p(\phi'|k')p(k')g'(u')}{p(X|\phi, k)p(\phi|k)p(k)g(u)} \left|\frac{\partial(\phi', u')}{\partial(\phi, u)}\right|\right\}, \tag{5}$$

where in our case, the absolute value of the Jacobian determinant is equal to one. We generate u_1, u_2, and u_3 independently, so that the proposal density factorizes as $g(u) = g_1(u_1)g_2(u_2)g_3(u_3)$. The reverse move requires only

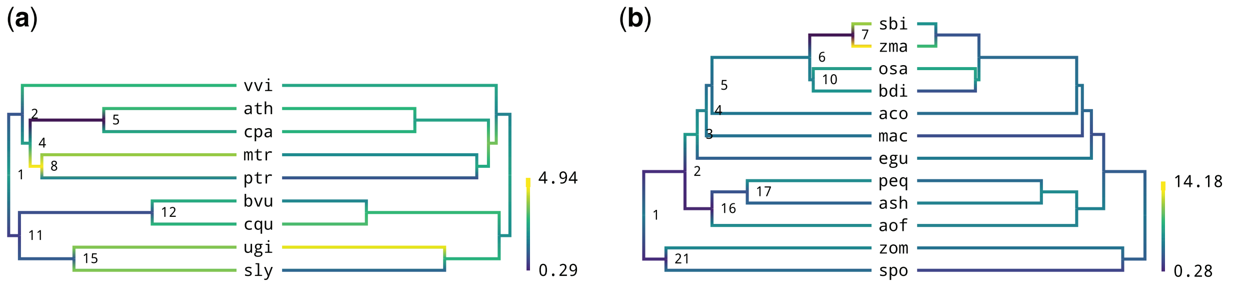

Fig. 4. Marginal posterior mean duplication (left tree in each panel) and loss rates (right tree in each panel) on a scale of $\log_{10}$ (events per lineage per billion year) for (a) the dicot data set ($\delta = 0.1$ analysis) and (b) the monocot data set ($\delta = 0.05$ analysis). Estimates correspond to the across-model geometric average. For species name abbreviations, we refer the reader to the Materials and Methods section.

one random variable u'_1, and we take $g_1(u_1) = g'_1(u'_1)$. We further sample u_3, that is, the WGD location along $\mathcal{S}$, from the Uniform$(0, T)$ prior. Collecting terms that appear in both the numerator and denominator and performing the appropriate cancellations, we obtain the acceptance probability of the forward move as

$$\alpha[(\phi,k),(\phi',k+1)] = \min\left\{1, \Lambda \frac{p(k+1)}{p(k)} \frac{p(\theta'_i|\theta_{-i},\Psi)}{p(\theta_i|\theta_{-i},\Psi)} \frac{p(q_j)}{g_2(q_j)}\right\}, \qquad (6)$$

where Λ denotes the likelihood ratio. Since we have a natural centering point in the sense of Brooks et al. (2003), that is, $\Lambda = 1$ if $u = u^* = 0$, we can see that under so-called weak nonidentifiability centering, an optimal choice for $g_2(u_2)$ may be to take the prior density $p(q)$.

We further implemented two slightly different reversible-jump kernels, 1) a kernel where not only λ_i but also μ_i is updated in a model jump and 2) a kernel where only a new WGD is proposed, without updating either λ_i or μ_i. The acceptance probability is identical to (6), and similar observations with regard to the choice of the proposal density of q_j hold. By default, we use the former of these two in our analyses. We verified the implementation of the MCMC algorithm by running the sampler in the absence of data, in which case the sample should approximately reproduce the prior (supplementary figs. S5 and S6, Supplementary Material online). Furthermore, we assessed our ability to recover accurate parameter estimates for data sets simulated under the DL and DLWGD models (see Results). Lastly, to assess the efficiency of the rjMCMC sampler for sampling across model space, we compute the effective sample size (ESS) for the number of WGDs (i.e., the model indicator k) using the method recently proposed by Heck et al. (2019).

Posterior Inference of WGD Configurations

In a Bayesian framework, model comparisons are usually performed using Bayes factors. The Bayes factor of a model M_1 versus a model M_0 is computed as

$$K_{10} = \frac{\Pr(X|M_1)}{\Pr(X|M_0)} = \frac{\Pr(M_1|X)/\Pr(M_0|X)}{\Pr(M_1)/\Pr(M_0)}.$$

This is straightforward to compute given an approximate sample of the posterior distribution across model space. Note however that we have treated WGD configurations with the same number of WGDs k (and as a result the same dimensionality of parameter space) as a single model M_k, yet model selection between different k is of limited interest. We are actually more interested in assessing whether some particular branch e is likely to be associated with a number k_e of WGDs. The approximate marginal posterior probability $p(k_e|X)$ of k_e WGDs on branch e is again easily obtained from the MCMC sample. Since the prior probability of a WGD on a particular branch e is proportional to the length of that branch t_e, the marginal prior probability $p(k_e)$ of k_e WGDs on branch e under, for instance, a Geometric(p) prior on the number of WGDs in $\mathcal{S}$ can be obtained as

$$\begin{aligned}
\Pr(k_e) &= \sum_{i=k_e}^{\infty} \Pr(k_e|k=i)\Pr(k=i) \\
&= \sum_{i=k_e}^{\infty} \binom{i}{k_e} \left(\frac{t_e}{T}\right)^{k_e} \left(1-\frac{t_e}{T}\right)^{i-k_e} (1-p)^i p \\
&= p\left(\frac{t_e}{T}(1-p)\right)^{k_e} \left[1-\left(1-\frac{t_e}{T}\right)(1-p)\right]^{-k_e-1}
\end{aligned}$$

For a Poisson(λ) prior on k, a similar argument shows that $k_e \sim \text{Poisson}(\lambda \frac{t_e}{T})$, whereas for any discrete prior with finite support on k, the first equality gives a means to compute the relevant prior probabilities. This enables us to compute Bayes factors for branch-specific WGD configurations by comparing a WGD configuration with k_e WGDs on branch e with a model with $k_e - 1$ WGDs on that branch. This will be our main strategy for employing samples acquired using the rjMCMC algorithm for inference of WGDs along $\mathcal{S}$.

Posterior Predictive Model Checks

After performing inference using the rjMCMC sampler, we use posterior predictive simulations to assess the fit of our inferred models to the data (Gelman et al. 2013; see Brown 2014; Höhna et al. 2018 for applications in phylogenetics). By evaluating whether data sets simulated from the posterior predictive distribution resemble the empirical data reasonably well, posterior predictive simulations serve to signal potential issues at the modeling level. We perform posterior predictive model checks by drawing 1,000 random models and associated parameters from the approximated joint-

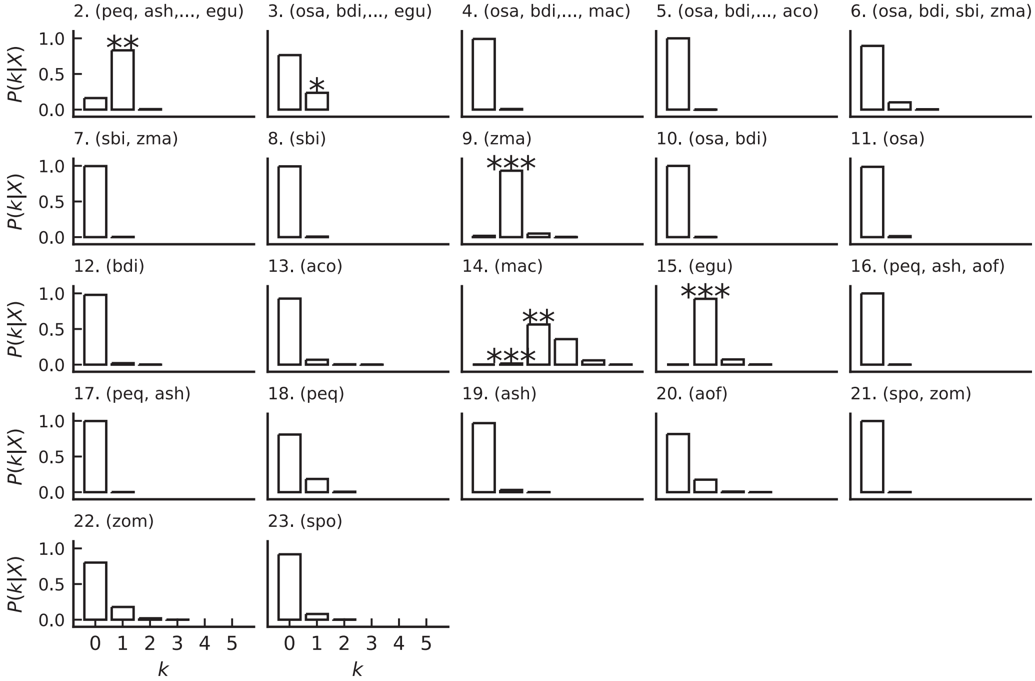

Fig. 5. Posterior probability of the number of WGDs on each branch in the monocot data set ($\delta = 0.05$ analysis). Interpretation is as in figure 3. For species name abbreviations, we refer the reader to the Materials and Methods section.

posterior distribution. For each such replicate we then simulate N phylogenetic profiles (i.e., a matrix with equal dimensions as the original data) from the associated DLWGD model. For these simulated data sets, we compute a series of summary statistics (mean, standard deviation, and entropy) for the numbers of genes at each leaf of $\mathcal{S}$ and the overall family size to acquire the posterior predictive distributions for these summary statistics. We then compare the observed values for these summary statistics with the relevant approximate posterior predictive distribution either graphically (e.g., supplementary figs. S22 and S27, Supplementary Material online) or numerically by computing posterior predictive P values (Gelman et al. 2013).

Gene Family Data

We obtained two data sets, one for a set of nine dicot species (*A. thaliana*, *Carica papaya*, *M. truncatula*, *P. trichocarpa*, *Vitis vinifera*, *U. gibba*, *C. quinoa*, *Beta vulgaris*, and *S. lycopersicum*) and another for a set of 12 monocot species (*Oryza sativa*, *Sorghum bicolor*, *Z. mays*, *Brachypodium dystachion*, *Ananas comosus*, *Elaeis guineensis*, *Musa acuminata*, *Asparagus officinalis*, *Phalaenopsis equestris*, *Apostasia shenzhenica*, *Spirodela polyrhiza*, and *Zostera marina*). In this article, in particular in figures, we refer to these species by three letter codes taking the first and first two letters from the genus and species names, respectively (except for *Zostera marina*, where we take "zom" to prevent collision with *Z. mays* ["zma"]). All sequence data were gathered from PLAZA dicots 4.0 and PLAZA Monocots 4.5 (Van Bel et al. 2018). The dated species tree for the dicots was based on a tree with median node ages from TimeTree (Kumar et al. 2017), whereas for the monocots, divergence times were gathered from Foster et al. (2016). Gene families were then obtained using OrthoFinder with default settings. We filtered out families that did not contain at least one gene in each clade stemming from the root of the associated species tree, to rule out de novo origination of gene families in arbitrary clades of the species tree. We condition all our analysis on this filtering procedure as in Rabier et al. (2014) and Zwaenepoel and Van de Peer (2019). We further filtered out large families using a Poisson outlier criterion, filtering out families for which $2Y > \text{median}(2Y) + 3$ where Y is the square root transformed family size. Unless stated otherwise, we restrict all our analyses to a random sample of 1,000 gene families from the filtered data sets for the sake of computational tractability.

Availability and Implementation

All methods were implemented in the Julia programing language (v1.3, Bezanson et al. 2017). Important computationally intensive routines support distributed computing in order to harness modern parallel computing environments. The associated software package is open source, documented, and freely available at https://github.com/arzwa/Beluga.jl (last accessed May 6, 2020). The data sets considered in this study are also available from that repository.

Results

Simulated Data

We conducted simulation studies to evaluate the statistical performance of the new methods we propose. All simulations were performed for the nine-dicot species tree. We first verified our ability to recover true parameter values for data sets of 500 gene families simulated from the independent rate (IR) prior under a reasonable yet broad range of parameter values. For this first set of simulations, we randomly sampled covariance matrices from the inverse Wishart prior with $\Psi = [0.5\,0.45;\ 0.45\,0.5]$ for 20 replicates. We then sampled branch-wise duplication and loss rates, assuming a multivariate lognormal prior with mean 1.5 and the same prior covariance matrix on the mean duplication and loss rates. These prior settings were based on preliminary analyses of a data set for the same species tree. For every replicate, we sampled a random number of WGDs with uniform retention rates, using a Geometric prior with $p = 0.2$ on the number of WGDs in the phylogeny. Values for η were sampled from a Beta$(5, 1)$ prior. We then performed posterior inference using fixed-dimensional MCMC assuming the simulated WGD configuration, with $\Psi = [1\,0.5;\ 0.5\,1]$, a prior mean duplication and loss rate of 1.5 and $\Sigma_0 = \Psi$. For each replicate, we sampled for 11,000 iterations, discarding 1,000 iterations as burn-in. Results for these simulations are shown in figure 1. We find that overall, even in the presence of considerable across-lineage rate variation, branch-wise duplication and loss rates can be estimated with fair accuracy and with no obvious biases across approximately three orders of magnitude. Retention rates are similarly well estimated for all but a handful of WGDs. In cases where multiple WGDs were present on the same branch, estimated retention rates were sometimes swapped (e.g., the inferred retention rate for the first simulated WGD corresponded to the retention rate for the second WGD on the same branch and vice versa), explaining the handful of seemingly incorrectly estimated retention rates (supplementary fig. S7, Supplementary Material online). We further note that in these simulations, an ESS $>$ 200 was obtained for all but a handful of parameters (supplementary fig. S8, Supplementary Material online).

Next, we studied the degree to which the assumption of constant rates across gene families affects our posterior inferences. For these simulations, we simulated for each replicate genome-wide lineage-specific rates θ from a multivariate normal distribution with a fixed covariance matrix with variance 0.1 and covariance 0.09. This allows for limited, though not negligible, rate heterogeneity across lineages. For each replicate, we sampled a series of 500 relative rates (r) from a Gamma$(\alpha, 1/\alpha)$ distribution with mean 1 and parameter α, with $\log_{10}\alpha \in \{-1, 0.5, 0, 0.5, 1\}$ multiplying for every gene family i the genome-wide rates by r_i. For every replicate, we thus obtained a set of 500 phylogenetic profiles from these family-specific BDPs. We then performed posterior inference under the IR model identical to our approach for the first set of simulations. These simulations indicated that for strongly asymmetric distributions of family-wise relative rates (small values of α), the genome-wide rates were systematically underestimated (supplementary figs. S9 and S10, Supplementary Material online). If the distribution of relative rates is not too extreme however, the genome-wide average lineage-specific duplication and loss rates were well recovered even though our inference is ignorant of across-family rate heterogeneity. It is relatively straightforward to include simple mixture models for across-family rate variation, such as Yang's discrete Gamma model (Yang 1994), however, this entails a considerable increase in computational time required for computing the likelihood of the data. As our models restricted to lineage-specific variation perform well if the distribution of rates across families is not too asymmetric, we do not further consider modeling family variation in this study.

To assess the performance of the reversible-jump MCMC algorithm for the inference of WGDs and branch-wise duplication and loss rates, we performed two sets of simulations with different branch-wise duplication and loss rates. For the first set of simulations, we simulated duplication and loss rates from the IR prior with a covariance matrix $\Psi = \Sigma_0 = [0.1\,0.05;\ 0.05\,0.1]$ and $\theta_0 = [\log(1.5)\ \log(1.5)]$, again allowing for limited across-lineage rate heterogeneity. For the second set of simulations, we first obtained a sample from the posterior distribution for the dicot data set using a fixed DLWGD model with well-known WGDs marked along the branches of the species tree (i.e., *Populus trichocarpa*, *Chenopodium quinoa*, *Arabidopsis thaliana* α and β, *Medicago truncatula*, *Solanum lycopersicum*, and three *Utricularia gibba* WGDs) (supplementary fig. S11, Supplementary Material online). For each simulation replicate, we then obtain a parameterized DLWGD model by drawing random branch-wise duplication and loss rates from the joint-posterior distribution acquired from the fixed-dimension MCMC sampler. Using this strategy, we ensure that the simulated duplication and loss rates are in a realistic range, allowing us to more properly assess the performance of the rjMCMC approach for WGD inference in reasonable settings, albeit tailored toward the dicot data set. In both sets of simulations, we randomly add six WGDs uniformly along the species tree with retention rates drawn from a Beta$(1.5, 2)$ distribution. For each simulation scheme, we simulated ten data sets of size $N = 100$, $N = 500$, and $N = 1{,}000$, respectively, to further assess the impact of using different amounts of data. We performed posterior inference using the default priors settings discussed in the Materials and Methods section, with a discrete uniform prior on the number of WGDs ranging from 0 to 10 WGDs (note that this will limit the number of possible false positives). We performed three sets of inferences with the IR prior using different parameterizations of the prior covariance matrix with $\Psi = \sigma^2 I_2$ and $\sigma^2 \in \{1, 0.2, 0.1\}$. For the second set of simulations, we also consider the effect of applying the constraining prior on the expected number genes per ancestral gene, where we assume a Normal$(1, 0.1)$ distribution on $\mathbb{E}[X_i|X_j = 1]$. We additionally perform inference using a constant-rates model where we assume the tree-wide duplication and loss rate are distributed as θ_1 in the IR model. All results are based on 10,000 iterations of the rjMCMC algorithm after discarding an initial 1,000 samples as burn-in.

We observe fair performance, and as expected, reliability tends to increase with the number of gene families, although this effect is not that strong (fig. 2). Similarly, the power to detect a WGD does not seem to increase when using more data. Decreasing the variance of the bivariate process from $\sigma^2 = 1$ to $\sigma^2 = 0.2$ or $\sigma^2 = 0.1$ does not seem to influence the number of false positive WGDs much for both sets of simulations, but smaller prior variances (allowing less rate heterogeneity) do seem to lead to increased false positive rates at least in the second simulation set. For the constant-rates model however, the number of false positive WGDs quickly rises to dramatic levels, and we note that here the prior on the number of WGDs in the species tree (a discrete uniform distribution from 0 to 10) prevents the number of false positive WGDs to rise even further. At the same time, we note for both sets of simulations that the power is not considerably higher when more restrictive models on the rate evolution process are assumed, at least for reasonable Bayes factor thresholds.

For the second set of simulations, a slight increase in the number of false positive inferences could be observed when the additional constraint prior was applied (i.e., the prior on $\mathbb{E}[X_i|X_j = 1]$). We note that this is not unexpected, as the simulated duplication and loss rates were themselves not estimated under a model including such a constraint. To assess whether the constraining prior improves inference when the data generating process is indeed closer to an equilibrium BDP (i.e., where $\mathbb{E}[X_i|X_j = 1] \approx 1$), we performed an additional set of simulations where we simulated data sets as in the first set of simulations, but assuming equal duplication and loss rates for each branch (but still with variable rates across branches). Here we observe a decreasing power to detect WGDs for increasing weakness of the constraint (i.e., increasing δ), in line with expectations supplementary figure S12, Supplementary Material online. These simulations show that in case the DL process is in equilibrium, with no net genome-wide expansion or contraction (which we deem a reasonable "neutral" expectation), applying this additional prior on the expected number of gene lineages per ancestral lineage may aid in WGD inference without having to assume an overly restrictive model of equal duplication and loss rates. At the same time, the simulations shown in figure 2 indicate that assuming such a prior when in fact some branches may not be in equilibrium does not lead to serious issues.

A main challenge in rjMCMC is to obtain decent mixing across model space in reasonable computational time (Rannala and Yang 2013). Besides rendering the likelihood evaluation computationally more demanding, using more data may also reduce the efficiency of the rjMCMC algorithm per se as the probability to jump between different WGD models becomes lower. We record that for the simulations from the IR prior it takes about 4 h to obtain 10,000 samples from the posterior on five CPUs for a data set of size $N = 1{,}000$ (supplementary fig. S13, Supplementary Material online) and the acceptance probability of a between-model move was about 0.20, 0.09, and 0.06 for the $N = 100$, $N = 500$, and $N = 1{,}000$ data sets, respectively, illustrating this phenomenon. This could cause the power not to increase with increasing data set size (as one would expect) due to a decrease in the efficiency of the rjMCMC sampler to explore model space. This is also clearly indicated by the ESS values obtained for the model indicator variable k (i.e., number of WGDs) across the simulated data sets, with the average ESS per 10,000 iterations decreasing with data set size (supplementary fig. S14, Supplementary Material online). Lastly, we observe that overall, marginal posterior distributions for duplication and loss rates tend to agree with the simulated values, however, for small data set sizes, branch-wise duplication rates are considerably "shrinked" toward the mean rate (supplementary fig. S15, Supplementary Material online). Investigating some replicates more closely (e.g., supplementary figs. S16 and S17, Supplementary Material online), we make two observations: 1) duplication and loss rates for a branch are generally harder to obtain accurately in the presence of WGD(s) on that particular branch and 2) duplication and loss rates for short internal branches tend to be biased toward the tree-wide mean rates (θ_0). We note that the latter observation reflects a feature that is not undesirable, because for short branches there will generally be less information in the data to accurately infer duplication and loss rates, and due to the hierarchical structure of our model these estimates tend to undergo "shrinkage" toward the tree-wide mean.

Plant Data Sets

We applied our new approach to two data sets from different plant taxa, focusing on well-sequenced genomes in dicots and monocots. These regions of the angiosperm phylogeny have been well-studied in terms of ancient WGDs, and, to some extent, a scientific consensus has emerged concerning which clades share particular ancestral WGD events, although considerable uncertainties remain (Jiao et al. 2014; Ming et al. 2015; Soltis and Soltis 2016; Van de Peer et al. 2017; Zhang et al. 2017). Based on some exploratory pilot runs, we use the IR prior with a prior covariance matrix $\Psi = 0.5I_2$, $\theta_0 = [\log 1.5, \log 1.5]$, $\Sigma_0 = [1.0, 0.9; 0.9, 1.0]$, $q \sim \text{Beta}(1, 3)$, and $\eta \sim \text{Beta}(3, 1)$, with the additional constraining prior on the expected number of lineages per ancestral lineage using $\delta = 0.1$ (unless stated otherwise). Throughout, we use a discrete uniform prior ranging from 0 to 20 for the number of WGDs. All results are based on at least 50,000 iterations of the rjMCMC algorithm and for all presented results the ESS associated with the number of WGDs was >500.

Applying the rjMCMC algorithm to the dicot data set (supplementary figs. S18 and S19, Supplementary Material online) revealed, as expected, strong support for WGD in poplar and quinoa, with in both cases posterior probabilities >0.9 for the one-WGD model (fig. 3). These WGDs are associated with very high retention rates, with a marginal posterior mean and 95% credibility interval $\bar{q} = 0.49\ (0.45, 0.53)$ for the poplar WGD event, and $\bar{q} = 0.90\ (0.85, 0.95)$ for the quinoa WGD. For *Arabidopsis*, we likewise find strong support for a WGD event, with the MAP model (at $P \approx 0.7$) the one-WGD model and a marginal posterior mean retention rate at $\bar{q} = 0.14\ (0.05, 0.20)$. We do not find strong support for a second WGD on this branch, indicating that gene count data may not contain sufficient signal to detect multiple WGDs on a single branch if these are associated with relatively low retention rates. We observe that for most branches, there is very strong evidence *against* WGD with the Bayes factor strongly favoring the no-WGD model. Interestingly, the two branches for which the Bayes factor is about one (and thus indecisive both ways), namely the branches leading to *Solanum* and *Medicago*, are branches associated with WGD.

The increased duplication rates for these branches suggest that the SSDL model is sufficiently flexible to accommodate the WGD-specific signal for long branches (fig. 4). Furthermore and perhaps unsurprisingly, we find a strong bimodal posterior distribution for the duplication rate on both branches, with the modes associated with the competing no-WGD and one-WGD models (supplementary fig. S20, Supplementary Material online). We further note that retention rates for the *Solanum* and *Medicago* events conditional on the one-WGD model for these branches were both estimated at $\bar{q} = 0.08\,(0.01, 0.15)$, which is virtually identical to the results obtained with a fixed-dimensional MCMC sampler, suggesting that our inability to find decisive support for these WGDs is not due to issues in the rjMCMC algorithm.

Assuming a stronger constraining prior on the expected number of lineages at the end of each branch, more specifically by setting $\delta = 0.05$, resulted in very strong support for these two WGDs (fig. 3), with posterior retention rates of $\bar{q} = 0.13\,(0.07, 0.17)$ and $\bar{q} = 0.14\,(0.07, 0.19)$ for the *Solanum* and *Medicago* WGDs, respectively. Posterior inferences for other branches are almost identical to the $\delta = 0.1$ analysis (supplementary figs. S18 and S19, Supplementary Material online), except for *A. thaliana*, where the posterior probability for the no-WGD model decreased to about zero (fig. 3), whereas the retention rate associated with the one-WGD model did not differ ($\bar{q} = 0.15\,(0.08, 0.19)$). This analysis again clearly illustrates how phylogenetic WGD inferences are strongly dependent on prior assumptions on the SSDL process. Lastly, we note that we do not find support for any of the WGDs in *U. gibba* in any of our analyses, whereas the consensus view is that this lineage has undergone three WGD events since its divergence from other Lamiales (Ibarra-Laclette et al. 2013); a view that was largely established based on comparative colinearity analyses. This lineage presents however a serious challenge to our approach, as its evolution has been associated with strong genomic reduction, characterized by, among others, extensive gene loss (Ibarra-Laclette et al. 2013; Carretero-Paulet et al. 2015). Our results suggest, in accord with this evolutionary history, that the purely quantitative signal from these WGDs has eroded considerably, and that they cannot be inferred from gene count data alone. We confirm the increased loss rate in this lineage, with an expected number of genes at the end of the branch per ancestral gene of $0.87\,(0.80, 0.92)$ (supplementary fig. S21, Supplementary Material online). Changing δ from 0.1 to 0.05 did not considerably alter this expected value, increasing it slightly to $0.89\,(0.84, 0.93)$. Overall, posterior predictive simulations indicate a fairly good model fit, with 25 out of 30 of the observed summary statistics in the 95% central mass of the empirical posterior predictive distributions (supplementary fig. S22, Supplementary Material online). In terms of the posterior predictive distributions, there were no noticeable differences between the chain with $\delta = 0.1$ and $\delta = 0.05$. As a point of reference, we note that for this small data set, we recorded a run time of about 2.2 h per 10,000 iterations for the dicot data set when employing ten CPUs.

We performed the same analyses for the monocot data set (figs. 4 and 5). We first performed the analysis with $\delta = 0.1$ but were unable to recover the well-known WGD in the maize lineage and observed several branches for which the expected number of lineages per ancestral lineage under the SSDL process alone was relatively high due to high estimated duplication rates (mainly in the banana [*Musa*] and maize [*Zea*] lineages, supplementary fig. S26, Supplementary Material online). An analysis with $\delta = 0.05$ resulted in an inferred SSDL process that was more or less in equilibrium and showed very strong support for the WGD in the maize lineage. We further report results for this second analysis with $\delta = 0.05$. We find strong support for the so-called τ WGD event shared by all monocots in our data set except the Alismatales (branch 2 in figs. 4 and 5), albeit with a fairly low retention rate $\bar{q} = 0.07\,(0.03, 0.10)$. As in Zhang et al. (2017), but unlike some other studies (Jiao et al. 2014; Ming et al. 2015), we find some support for a shared WGD event in the ancestor of all commelinids (branch 3), which we suggest to be the event referred to by σ, generally thought to be shared by all Poales (Ming et al. 2015). We note however that trace plots indicate that mixing across model space seems to be challenging for this branch (supplementary fig. S23, Supplementary Material online), resulting in rather uncertain estimates for the posterior probability of this WGD (the method of Heck et al. [2019] suggests a 95% posterior interval of [0.10, 0.54] for the posterior probability of the one-WGD model for this branch). Again, this putative WGD event is associated with a relatively low retention rate $\bar{q} = 0.05\,(0.02, 0.08)$. We find no support for the cereal-specific genome duplication event referred to by ρ (fig. 5, branch 6) in our gene count data, and find that the SSDL process along this branch is more or less in equilibrium, with the duplication rate ($\bar{\lambda} = 1.40\,(0.71,\ 1.75)$) only slightly higher than the loss rate ($\bar{\mu} = 1.10\,(0.83,\ 1.36)$) along this branch (fig. 4 and supplementary fig. S26, Supplementary Material online). This is unlike our observations for the *Medicago* and *Solanum* branches in the nine-taxon dicot data set, where the absence of decisive support for the WGD events in these lineages was associated with increased duplication rates. The results for these putative ancestral WGD events in the monocot data set could indicate that the power of the gene count based rjMCMC approach for detection of ancestral WGDs might be affected by taxon-sampling issues, where WGDs on short branches leading to a species rich crown group can be detected more easily.

As already indicated, we find very strong support for WGD in the maize (*Zea*, $\bar{q} = 0.15\,(0.09, 0.19)$) lineage when the prior with $\delta = 0.05$ is used. The difficulty associated with identifying this WGD has likely to do with the high duplication rate in this lineage (fig. 4), which is notably higher than the rate in all other lineages in the monocot phylogeny considered here. We further find overwhelming support for WGD in the oil palm (*Elaeis*, $\bar{q} = 0.28\,[0.23, 0.33]$) lineage, in line with previous results (Singh et al. 2013; Jiao et al. 2014). We also recovered the well-known multiple events along the branch leading to *Musa* (D'Hont et al. 2012), with strong support for a two-WGD model in the $\delta = 0.1$ analysis and a three-WGD model in the analysis with $\delta = 0.05$. The different WGDs can however not be distinguished, as exemplified by the fully overlapping distributions of the retention rates for the WGDs when ordered by the WGD time (supplementary fig. S25, Supplementary Material online). Again, we find very strong evidence *against* WGD for most other branches, notably the branch leading to the orchids where a WGD was hypothesized by Zhang et al. (2017). The only branches that have some posterior probability for a one-WGD model (apart from the already discussed ρ WGD) are those leading to *Asparagus*, *Phalaenopsis*, *Zostera*, and *Spirodela*. These lineages are thought to have underwent ancient WGDs (Wang et al. 2014; Cai et al. 2015; Olsen et al. 2016; Harkess et al. 2017) and are, in our analyses, suggestively, associated with gene family expansion (i.e., a nonequilibrium SSDL process supplementary fig. S26, Supplementary Material online). However, with the taxon sampling employed here, we find that gene count data alone do not provide enough signal to distinguish a flexible SSDL model from a DLWGD model. Posterior predictive model simulations indicate a similarly fair fit as for the dicots data set (supplementary fig. S27, Supplementary Material online), with 32 and 29 out of 39 summary statistics within the 95% central mass of the marginal posterior predictive distribution for the analysis with $\delta = 0.05$ and $\delta = 0.1$, respectively.

Discussion

In this article, we continued our previous work on the statistical inference of ancient WGD events from comparative genomic data (Zwaenepoel and Van de Peer 2019). Building on the initial idea of Rabier et al. (2014), we model gene family evolution by SSDL using a phylogenetic linear BDP and exploit the statistical deviation from such a process to infer ancient WGDs. In contrast to the approaches taken in these previous studies, we here do not assume a particular WGD configuration and infer the number and locations of WGDs in a species tree from the data using reversible-jump MCMC under a flexible hierarchical model of gene family evolution. In our previous work, we showed that using simple birth–death models for the SSDL process, in particular models assuming constant rates of duplication and loss across the entire phylogeny, seriously compromises reliable inference of ancient WGDs. Simulations showed that false positive rates for WGD inference become unacceptably high when the constant-rates assumption is violated (Zwaenepoel and Van de Peer 2019), which we deem likely to be the rule rather than the exception for most phylogenies. On the other hand, employing more complicated, hierarchical models with branch-wise duplication and loss rates may compromise our ability to detect WGDs from gene count or gene tree information, with the assumed SSDL process sufficiently flexible to capture a WGD signature. In the Bayesian inference scheme, this tradeoff translates to a certain degree of prior sensitivity, with assumptions on the rate evolution process potentially affecting our posterior inferences.

In this work, we are confronted with the very same challenges. Using a flexible model for the SSDL process is vital to prevent false positive WGD inferences but concomitantly results in a statistical approach with limited power to detect true WGD events for long branches or WGD events with small associated retention rates. In our multivariate models of duplication and loss rate evolution, this flexibility is embodied by the covariance matrix of the branch-wise duplication and loss rates, which affects both the amount of rate heterogeneity across branches and the difference between the duplication and loss rate for any particular branch. We further optionally constrain the latter by using an additional prior on the expected number of lineages per ancestral lineage under the SSDL process, which has the advantage of being more intelligible than the prior covariance parameter, therefore allowing easier application of informative priors. Nevertheless, even when using fairly informative priors in the latter form, we find that information from gene count data often does not provide decisive support for putative WGD events previously described in the literature.

These statistical problems of power and prior sensitivity are in the first place due to the kind of data employed for tackling the problem of WGD inference. Indeed, the same issues are relevant in some form or another whenever inference is based on phylogenomic data—whether in the form of gene trees of multicopy gene families (as in Zwaenepoel and Van de Peer [2019], Li et al. [2018], 1KP initiative [2019], etc.) or gene counts (as in Rabier et al. [2014] and this study)—and ignores information from genome structure. By using a rich model, we seek to explicitly account for the most relevant sources of variation in this data when assessing the history of WGDs in a phylogeny, and we are able to assess the effects of particular assumptions of the SSDL process on our WGD inferences. We believe the Bayesian approach advocated here compels us to embrace the inherent uncertainty of our inferences based on this kind of data, while allowing WGD inference in a coherent and biologically meaningful framework. The usage of gene trees instead of gene counts may further increase the power for WGD detection (Zwaenepoel and Van de Peer 2019), and the accuracy of duplication and loss rate estimates. However, this comes at the expense of a less efficient algorithm for computing the likelihood and a much more involved and computationally intensive data preparation phase (which in the Whale approach of Zwaenepoel and Van de Peer [2019] involves obtaining a sample from the posterior distribution of gene tree topologies for every gene family). Using gene counts instead could allow us to employ broader sets of taxa, which may be beneficial, as improved taxon sampling could result in a higher power to detect ancient WGDs, although this has to be studied in more detail. To us, the most fruitful avenue for future research appears to involve an integration of information from genome structure in the probabilistic phylogenomic framework we adopt here; leveraging the obvious differences between the SSDL and WGD process at the level of gene synteny or colinearity. Lastly, we note that although posterior predictive simulations suggest that the DLWGD model with branch-wise rates provides a reasonable fit to the data, there may be room for improvement on the modeling side as well. However, more complicated models (e.g., involving nonlinear BDPs [Crawford et al. 2014] or rate heterogeneity across families) will generally be associated with even higher computational demands. Another potentially interesting improvement in terms of modeling would be to account for incompletely sampled genomes and assembly or annotation errors (as, e.g., in Han et al. [2013] or Rabier et al. [2014]) in the Bayesian approach we adopt here.

Computational problems are however more of an issue than the statistical intricacies discussed above, which are after all reflecting the very evolutionary processes of interest. Whereas we have implemented the likelihood evaluation routines and rjMCMC algorithm in an (to the best of our knowledge) efficient manner and exploit distributed computing architectures to accelerate computation of the likelihood across families; we nevertheless had to restrict our analyses to relatively limited sets of taxa and subsets of the full data. Furthermore, whereas mixing within a particular WGD configuration tends to be very efficient irrespective of the data set size, larger data sets tend to be associated with poor mixing across model space in the rjMCMC, with the acceptance probability for trans-dimensional moves dropping rapidly with increasing data set size. This confers a risk that for larger data sets an incomplete view of model uncertainty is obtained, and it forces us to sample longer chains to ensure decent mixing of the chain across model space; making the problem even more computationally intractable. Instead of investing computer power in the analysis of more gene families, it may be worthwhile to instead focus on adding more taxa, as this could similarly improve the statistical power for WGD inference while not affecting mixing efficiency. We note that as a follow-up analysis, a resulting set of putative WGDs could further be tested extensively using a fixed-dimension MCMC sampler in a manner analogous to Rabier et al. (2014) and Zwaenepoel and Van de Peer (2019), possibly with a larger data set. This could result in a more accurate approximation of the posterior distribution of the duplication, loss and retention rates under a particular DLWGD model compared with the approximated distribution from the rjMCMC sample. Finally, we have in this study only considered WGD, whereas at least some ancestral polyploidization events are associated with higher multiplication levels such as the γ triplication in eudicots (Jiao et al. 2012) and the Solanaceae triplication (Tomato Genome Consortium 2012). It is possible to extend the DLWGD model to higher multiplication levels (Rabier et al. 2014), and in the Bayesian approach we adopt here, the multiplication level could in principle be included as an additional parameter, or alternatively, reversible-jump moves for different multiplication levels could be included. We did not however explore this additional layer of complexity in our current study and defer this to future work.

In summary, we present a fully model-based approach for WGD detection in a phylogenetic context using a flexible hierarchical model of gene family evolution. Posterior inference is based on trans-dimensional MCMC using the reversible-jump Metropolis–Hastings algorithm. We model variation of duplication and loss rates using a multivariate approach, and we note that this in principle would allow to model correlated evolution with other quantitative traits as in Lartillot and Poujol (2011), although we did not explore this in our study and defer this to future research efforts. Through simulations and analyses of comparative genomic data from flowering plants, we investigated the reliability and power of our new approach and provide further insights in the statistical inference problem of detecting ancient WGDs from phylogenomic data. We find that when a flexible model of gene family evolution is assumed, the power for WGD inference is rather limited, with strong support across all analyses only for those WGDs with a very strong signature (e.g., the WGDs in the *C. quinoa*, *P. trichocarpa*, *A. thaliana*, *Musa acuminata*, and *Elaeis guineensis* lineages in our analyses). For some other putative WGDs (*M. truncatula*, *S. lycopersicum*, and *Zea mays*), we find that our inferences are sensitive to prior assumptions, indicating that some caution is warranted when applying the proposed methods. In general, we believe performing multiple analyses under progressively more restrictive priors may provide insights in which WGD hypotheses are supported under which assumptions on the background SSDL process. Computationally, our rjMCMC approach is challenging, especially for large data sets, and the Bayesian inference machinery underlying this work is an obvious target for future improvements. Eventually, we hope that the methods presented here enable statistically better informed inferences of ancient WGDs and can contribute to an improved understanding of this key and increasingly appreciated process in genome evolution.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank two anonymous reviewers for thoughtful comments and suggestions that improved the manuscript. A.Z. is supported by a PhD fellowship of the Research Foundation Flanders (FWO). Y.V.d.P. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (Grant Agreement No. 833522).

References

1KP initiative. 2019. One thousand plant transcriptomes and phylogenomics of green plants. *Nature* 574:679–685.

Bezanson J, Edelman A, Karpinski S, Shah VB. 2017. Julia: a fresh approach to numerical computing. *SIAM Rev*. 59(1):65–98.

Brooks SP, Giudici P, Roberts GO. 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *J R Statist Soc B* 65(1):3–39.

Brown JM. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Syst Biol*. 63(3):334–348.

Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Chen LJ, He Y, Xu Q, Bian C, et al. 2015. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet*. 47(1):65–72.

Carretero-Paulet L, Librado P, Chang TH, Ibarra-Laclette E, Herrera-Estrella L, Rozas J, Albert VA. 2015. High gene family turnover rates and gene space adaptation in the compact genome of the carnivorous plant *Utricularia gibba*. *Mol Biol Evol*. 32(5):1284–1295.

Crawford FW, Minin VN, Suchard MA. 2014. Estimation for general birth–death processes. *J Am Stat Assoc*. 109(506):730–747.

Csűrös M, Miklós I. 2009. Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol*. 26(9):2087–2095.

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488(7410):213–217.

Foster CS, Sauquet H, Van der Merwe M, McPherson H, Rossetto M, Ho SY. 2016. Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst Biol*. 66:338–351.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. Bayesian data analysis. London: Chapman and Hall/CRC.

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82(4):711–732.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*. 15(8):1153–1160.

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using cafe 3. *Mol Biol Evol*. 30(8):1987–1997.

Harkess A, Zhou J, Xu C, Bowers JE, Van der Hulst R, Ayyampalayam S, Mercati F, Riccardi P, McKain MR, Kakrana A, et al. 2017. The *Asparagus* genome sheds light on the origin and evolution of a young y chromosome. *Nat Commun*. 8(1):1279.

Heck DW, Overstall AM, Gronau QF, Wagenmakers EJ. 2019. Quantifying uncertainty in transdimensional Markov chain Monte Carlo using discrete Markov models. *Stat Comput*. 29(4):631–643.

Höhna S, Coghill LM, Mount GG, Thomson RC, Brown JM. 2018. P3: phylogenetic posterior prediction in RevBayes. *Mol Biol Evol*. 35(4):1028–1034.

Huelsenbeck JP, Larget B, Swofford D. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154(4):1879–1892.

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juárez MJA, Simpson J, et al. 2013. Architecture and evolution of a minute plant genome. *Nature* 498(7452):94–98.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol*. 13(1):R3.

Jiao Y, Li J, Tang H, Paterson AH. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* 26(7):2792–2802.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 34(7):1812–1819.

Lartillot N, Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol*. 28(1):729–744.

Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A*. 115(18):4713–4718.

Li Z, Tiley GP, Rundell RJ, Barker MS. 2019. Reply to Nakatani and McLysaght: analyzing deep duplication events. *Proc Natl Acad Sci U S A*. 116(6):1819–1820.

Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* 28(2):279–281.

Liu L, Yu L, Kalavacharla V, Liu Z. 2011. A Bayesian model for gene family evolution. *BMC Bioinformatics* 12(1):426.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet*. 47(1):307–333.

Lynch M. 2007. The origins of genome architecture. Vol. 98. Sunderland (MA): Sinauer Associates.

Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E, et al. 2015. The pineapple genome and the evolution of cam photosynthesis. *Nat Genet*. 47(12): 1435–1442.

Muller HJ. 1936. Bar duplication. *Science* 83(2161):528–530.

Nakatani Y, McLysaght A. 2019. Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proc Natl Acad Sci U S A*. 116(6):1816–1818.

Novozhilov AS, Karev GP, Koonin EV. 2006. Biological applications of the theory of birth-and-death processes. *Briefings Bioinf*. 7(1):70–85.

Olsen JL, Rouzé P, Verhelst B, Lin YC, Bayer T, Collen J, Dattolo E, De Paoli E, Dittami S, Maumus F, et al. 2016. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature* 530(7590):331–335.

Rabier CE, Ta T, Ané C. 2014. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol Biol Evol*. 31(3):750–762.

Rannala B, Yang Z. 2013. Improved reversible jump algorithms for Bayesian species delimitation. *Genetics* 194(1):245–253.

Roberts GO, Rosenthal JS. 2009. Examples of adaptive MCMC. *J Comput Graph Stat*. 18(2):349–367.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM. 2018. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol*. 2(5):890–896.

Singh R, Ong-Abdullah M, Low ETL, Manaf MAA, Rosli R, Nookiah R, Ooi LCL, Ooi SE, Chan KL, Halim MA, et al. 2013. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature* 500(7462):335–339.

Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Curr Opin Plant Biol*. 30:159–165.

Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S. 2017. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *Plant Cell* 29(11):2766–2785.

Tiley GP, Ane C, Burleigh JG. 2016. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biol Evol*. 8(4):1023–1037.

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635.

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K. 2018. Plaza 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res*. 46(D1):D1190–D1196.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet*. 18(7):411–424.

Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo M, Lomsadze A, Borodovsky M, Kerstetter R, Shanklin J, et al. 2014. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat Commun*. 5(1):3311.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39(3):306–314.

Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, Niu SC, Wang JY, Lin YC, Xu Q, Chen LJ, et al. 2017. The *Apostasia* genome and the evolution of orchids. *Nature* 549(7672):379–383.

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat Ecol Evol*. 3(4):679–690.

Zwaenepoel A, Li Z, Lohaus R, Van de Peer Y. 2019. Finding evidence for whole genome duplications: a reappraisal. *Mol Plant* 12(2):133–136.

Zwaenepoel A, Van de Peer Y. 2019. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol Biol Evol*. 36(7):1384–1404.